APPENDIX C-4
Record Content Analysis Methodology

APPENDIX C-4
Record Content Analysis Methodology

Table of Contents

1.0. Introduction 1
2.0. Method Overview 1
3.0. Objectives 1
4.0. Context Within The Evaluation Framework 2
5.0. Data Collection and Analysis 2
5.1. Survey of GILS Universe 2
5.2. Development of Analysis Criteria 4
5.2.1. Issues in Developing Record Content Aggregation Criteria 5
5.3. Content Analysis of Sampled Records 6
6.0. Method Limitations and Recommendations to Future Researchers 8
7.0. Conclusion 8

Table C4-1 Record Content Analysis Sample Population 3
Table C4-2 Record Content Analysis Criteria 4
Table C4-3 Aggregation Semantics 6
Table C4-4 Information Object Semantics 7

APPENDIX C-4
RECORD CONTENT ANALYSIS METHODOLOGY

1.0. INTRODUCTION

Moen and McClure, in *The Government Information Locator Service (GILS): Expanding Research and Development on the ANSI/NISO Z39.50 Information Retrieval Standard: Final Report* (1994, p. 30) noted "an important factor in the overall utility of the GILS will be the quality of the data in GILS records. Quality criteria will include accuracy, consistency, completeness, and currency. In order to encourage the creation of high quality information that will populate GILS servers, the development of written guidelines for creating GILS records is essential." This direction, *The government information locator service: Guidelines for the Preparation of GILS Core Entries* (National Archives and Record Administration, 1995a) is available electronically from the National Archives gopher at <gopher.nara.gov> under "Information for Archivists and Records Managers/GILS Guidance," or from <URL: http://www.nara.gov:70/1/managers/gils>. In addition, *Federal information processing standards publication 192, Application Profile for the Government Information Locator Service (GILS)* (National Institute for Standards and Technology, 1994) provides other quality-related direction such as preferred order of display for record elements as well as their definitions.

Content analysis of GILS records served three purposes: to assess records' quality in terms of completeness and accuracy; to explore the relationship of selected characteristics of records and serviceability in networked information discovery and retrieval (NIDR); and to develop recommendations for future application or adaptation of the method.

More than 3500 instances of metadata were evaluated for incidence and/or content, and entered into a database for coding and analysis. In addition, the evaluators maintained a log of lessons learned and areas for further research (see Appendix E-2 Record Content Analysis Findings, Discussion, and Recommendations) that may be utilized by system developers, specification and procedures writers, and people with direct responsibility for GILS record quality.

2.0. METHOD OVERVIEW

The analysis comprised in two phases: Phase 1 involved examination of a pool of 83 records from 42 agencies' GILS retrieved deliberately to represent a range of information resource types (e.g., databases, catalogs, records systems). These records served as the basis for developing and operationalizing a set of more than 50 qualitative and quantitative evaluative criteria that included records' format, aggregation, media representation, and descriptiveness. Descriptiveness was defined as the incidence of utilization and content (value) attributes for all mandatory and selected optional elements and subelements as specified by *FIPS Pub. 192* Annex E-GILS Core Elements and the NARA *Guidelines*. In Phase 2, these criteria were systematically applied to a set of 83 records randomly retrieved January 13 and 14, 1997, from 42 agencies' GILS.

The following paragraphs present information concerning the record content analysis objectives, the context of the analysis within the overall evaluation framework, data collection and analysis, method limitations, lessons learned, and recommendations.

3.0. OBJECTIVES

This analysis attempted to describe the "quality" of GILS records in terms of character or attributes rather than strict conformance to specifications. The latter, which constitutes an audit, would require a greater level of operational detail than current policy and standards provide and is a technique better suited to a more mature information service. The following objectives guided the current examination of GILS records. Where adherence to published direction was relevant, *FIPS Pub. 192* Annex E definitions, as reproduced and supplemented by usage guidelines and examples in the NARA *Guidelines*, served as the basis for evaluation:

1. To assess the accuracy of GILS records in terms of errors in format and spelling
2. To gauge and compare the relative "completeness" or level of description of GILS records
   - Number of elements per record ("blank" vs. populated)

- Utilization and values of both mandatory and selected optional elements

3. To characterize a general profile of GILS product in terms of record types, aggregation levels, and containers (dissemination media)
4. To evaluate records' serviceability
   - Factors affecting NIDR
   - User convenience
   - Aesthetics and readability
   - Relevance judgment.

The quantitative and qualitative assessments, respectively, of the constitution and properties of sampled records provided data meeting these objectives.

4.0. CONTEXT WITHIN THE EVALUATION FRAMEWORK

As with the other methods comprising this user-oriented evaluation of GILS implementations, the record content analysis both was informed by and served to inform other data collection and instrument development activities in the study. Presentations and panel discussions at the 1996 GILS Conference and focus groups with various user communities highlighted recurring issues surrounding the content of GILS records, such as the level of resource aggregation, suitability of metadata elements, consistency, and quality of presentation. In turn, as discussed in Appendix E-2 Record Content Analysis Findings, Discussion, and Recommendations, the record content analysis proved invaluable in developing a user-assessment script that would both isolate GILS "quality" from that of the user interface or search engine and present realistic information retrieval encounters.

5.0. DATA COLLECTION AND ANALYSIS

Data collection and analysis were performed as described in the following paragraphs using the tool presented in Appendix D-4 Record Content Analysis Instrument as constructed in a Microsoft Access© database and Microsoft Excel© spreadsheets. Two surveying activities were prerequisite to the analysis of record content: a determination of the GILS universe to optimize the breadth of the sample and a review of planned (i.e., per the NARA *Guidelines*) vs. actual record characteristics to inform development of analysis criteria.

5.1. Survey of GILS Universe

To provide the broadest possible base for record selection, the investigators first determined the universe of GILS implementations. This was accomplished through various means:

- Verbal or written mention during the 1996 GILS Conference presentations and in handouts and survey responses, respectively
- Linking from the White House website's "President's Cabinet" <http://www.whitehouse.gov/WH/Cabinet/html/cabinet_links-plain.html> and "Federal Agencies and Commissions" <http://www.whitehouse.gov/WH/Independent_Agencies/html/independent_links-plain.html> to agency homepages, which, in turn, linked in some cases to FedWorld GILS <http://fedworld.gov/gils>
- WWW searches by means of Alta Vista <http://www.altavisa.digital.com> and Lycos <http://www.lycos.com> search engines for Executive department and agency names
  - As delineated in the *1996-97 Government Manual* via the Government Printing Office (GPO) GPO Access http://www.access.gpo.gov/su_docs/aces/aaces002.html>
  - As comprising the Chief Information Officer Council as specified in Executive Order 13011 of July 16, 1996 "Federal Information Technology" (http://www.gsa.gov/irms/ka/regs/exo13011/exo13011.htm)
- WWW searches by means of Alta Vista and Lycos search engines for "GILS" and for "government information locator service"
- GPO Access GILS server.

Results of this effort, completed on December 31, 1996, are shown in below in Table C4-1 Record Content Analysis Sample Population with two additional agencies identified for sampling in Phase 2 of the record content analysis.

Table C4-1
Record Content Analysis Sample Population

Consumer Product Safety Commission
Department Of Agriculture
Department Of Commerce
Department Of Defense
Department Of Energy
Department Of Health And Human Services
Department Of Housing And Urban Development
Department Of Interior
Department Of Labor
Department Of State
Department Of Treasury
Environmental Protection Agency
Equal Employment Opportunity Commission
Farm Credit Administration
Federal Communications Commission
Federal Emergency Management Agency
Federal Energy Regulatory Commission
Federal Labor Relations Authority
Federal Maritime Commission
Federal Reserve Board
Federal Trade Commission
General Services Administration
Government Printing Office
International Trade Commission
Merit Systems Protection Board
National Aeronautics And Space Administration
National Archives And Records Administration
National Transportation Safety Board
Nuclear Regulatory Commission
Nuclear Waste Technical Review Board
Office Of Government Ethics
Office Of Management And Budget
Office Of Personnel Management
Overseas Private Investment Corporation
Pension Benefit Garanty Corporation
Railroad Retirement Board
Securities And Exchange Commission
Selective Service System
Small Business Administration
Social Security Administration
U.S. Commission On Civil Rights
U.S. Postal Service
Total=42

5.2. Development of Analysis Criteria

The second activity to prepare for a systematic analysis of GILS record content was the creation of criteria to satisfy the study objectives. This was accomplished by examining a set of two records retrieved from each identified agency GILS. These records—retrieved by use of search terms including "system," "database," "manual," the agency acronym, subject-oriented single words—were selected to represent a variety of file sizes, formats, and content types.

These records were studied and compared to produce the assessment categories shown in Table C4-2 Record Content Analysis Criteria. (Appendix D-4 Record Content Analysis Instrument presents a table of the database fields, possible values, and coding notes that was constructed to record data.)

Table C4-2
Record Content Analysis Criteria

Accuracy

- Format and Formatting Errors
- Spelling And Typographical Errors

Completeness

- Number Of Elements Per Record
- Practice Of Presenting "Blank" (Nonpopulated But Displayed) Elements
- Utilization And Selected Characteristics Of "Mandatory" Elements
  - Title
  - Originator
  - Local Subject Index
  - Abstract
  - Purpose
  - Agency Program
  - Availability-Distributor
  - Availability-Order Process
  - Sources Of Data
  - Access Constraints
  - Use Constraints
  - Point Of Contact
  - Schedule Number
  - Control Identifier
  - Record Source
  - Date Of Last Modification
- Utilization And Characteristics Of Selected "Optional" Elements
  - Controlled Vocabulary-Index Terms-Controlled
  - Controlled Vocabulary-Thesaurus
  - Local Subject Index
  - Availability-Resource Description
  - Methodology

Profile

- Record Types (AIS, locator, Privacy Act system)
- Record Aggregation (See Table C4-3 Aggregation Semantics and discussion)
- Objects Represented (see Table C4-4 Information Object Semantics)
- Containers (Dissemination Media)
  - Broadcast (Radio/TV)
  - CD-ROM
  - Dialup
  - Email
  - Fax
  - Ftp Site
  - Gopher Site
  - Listserv
  - Microform
  - Multiple
  - Print
  - Video
  - Voice
  - Web

Serviceability

- Capitalization
- Citation Of Legislation
- Definition Of Acronyms
- Element Display Order
- Fielded-Search Option
- File Formats
- Hypertext
- Indentation
- Locally Defined Elements

5.2.1. Issues in Developing Record Content Aggregation Criteria

The following definitions served as an initial starting point for operationalizing the phenomenon of aggregation:

> **AGGREGATION**: the degree to which two or more separate parts have been brought together without changing their function or producing any result other than the sum of the operation of the parts.
>
> **GRANULATION**: the degree to which two or more separate parts of a whole are distinguishable within that whole.

It became apparent during review of the Phase 1 sample that the above definitions are unsuitable for application to GILS records. For example, a record describing a publicly-accessible enterprise-wide AIS whose function is to track information output of four discrete, functionally dedicated, not publicly accessible micro-AISs could be labeled a "highly aggregated" record in that it "rolls up" other potential records. But, should the record include a description of each "grain" (microsystem) it embraces, one would be tempted to code it "low granularity" (subparts are distinguishable).

Another, more concrete, example of the problem of characterizing aggregation of information resources would be *The Federal Register* in digital (databased) or paper print format. This one record describes one "discrete" publication, but that publication aggregates myriad standalone information objects that, in print, are highly granular to the initiated user but in database form (digital format) are less distinguishable.

Another, more concrete, example of the problem of characterizing aggregation of information resources would be *The Federal Register* in digital (databased) or paper print format. This one record describes one "discrete" publication, but that publication aggregates myriad standalone information objects that, in print, are highly granular to the initiated user but in database form (digital format) are less distinguishable.

In short, the attribute of "aggregation" is discernible only to the degree that the GILS record presents an explicit enumeration of "granules" or aggregated parts—whether those parts are:

- book chapters,
- database fields,
- Web page titles, or
- Privacy Act records,

which some will argue is too granular, or they are:

- individual reporting systems of enterprise-wide AIS,
- titles within a videotape series, or
- memoranda within a "file,"

which some will argue should be distinguishable.

Application of definitions of aggregation and granularity imply a knowledge of component-level and collective functionalities that the investigators, and, by proxy, a GILS user, lack and which may be gained only through examination of the object. In a physical library, users of a card catalog, subject bibliography, or other metadata-based tools are accustomed to retrieving and scanning resources' object-peculiar "primary" metadata (e.g., tables of content, graphics, and back-of-the-book indexes) as required to determine whether "granules" might satisfy their information need; in GILS, where often information resources cannot be examined and thus their "operation" is unknown, the concept of simply "pointing" to an aggregated "locator" may be insufficient in that the aggregation "produces no result other than the sum of the operation of the parts."

Nonetheless, because record and resource aggregation was identified as a recurring theme during other data collection activities of the study, investigator's adopted the operational definitions of aggregation coding scheme shown in Table C4-3 Aggregation Semantics to characterize the phenomenon. To supplement the limited value returned from assigning aggregation-level coding, investigators incorporated the criterion of "information object" as defined in Table C4-4 as well. Appendix E-2 Record Content Analysis Findings, Discussion, and Recommendations offers additional interpretation of the utility of these measures relative to aggregation and resource description.

Table C4-3
Aggregation Semantics

Code	Operational Definition	Examples
Record Aggregates Objects	GILS record, by virtue of its creation, collects discrete information resources that record content indicates would not have otherwise been collected or aggregated. Assigned in the absence of clues within the record that the represented objects were heretofore packaged *as this collection* to optimize information discovery and retrieval.	• Privacy Act Systems compilation • files • press releases • forms
Aggregated Object Represented	GILS record represents an *a priori* or purposeful collection of information resources—e.g., woodpecker database or agency website. GILS record represents an object that collects, or comprises, two or more discrete information objects, and that represents a collection of standalone information files or products packaged together on the basis of a common theme or subject for functional convenience.	• CD-ROM of regulations • System that compiles Privacy Act records • job line of open requisitions
Discrete Object Represented	GILS record describes a standalone document-level entity that does not meet the criteria for "object aggregates metadata" below.	• annual report • videotape
Object Aggregates Metadata	GILS record describes a pre-existing metadata collection, or "locator," as an information resource.	• directory • catalog • index • log

5.3. Content Analysis of Sampled Records

As of early January 1997, 42 agencies' GILS had been discovered by procedures identified in Section 5.2 Survey of the GILS Universe. The 83 sampled records, selected as described in the next paragraph, resided in three broad "host" categories: GPO (61% of the sample), record sources (34%), and FedWorld (5%). 93% of sampled records resided on a WAIS or Z39.50-compliant server, with the remaining on an HTTP server containing standalone HTML files of GILS records. (Note: since the time period of analysis, FedWorld and GPO have mounted record-source hosted GILS and those hosted by one another, and at least one HTTP-based GILS has migrated to WAIS).

The record content analysis *per se* first involved selection of GILS records from the known GILS universe (see Table C4-1 Record Content Analysis Sample Population) in one of two ways. For GILS featuring a search engine (i.e., residing on an information retrieval-based platform such as WAIS or Z39.50-compliant server or including a site-resident search engine), the investigator retrieved the first and last "hits" resulting from a "full-text" query of the agency acronym (using the default "number of records to return"). For GILS on which this was not possible (i.e., those mounted on a web server of HTML files that present only a picklist of record titles as if for known-item retrieval or browsing), the investigator retrieved the first and last items listed. In the event of multiple record formats per record, the HTML format was selected.

The resultant 83 records (one agency's GILS featured only 1 record) were printed for ease of study and comparative reference. Their characteristics were assessed and recorded in a relational database for compilation and subsequently transferred to a spreadsheet for analysis using descriptive statistics. A subset of the total was created and subject to identical analysis by filtering the data for values of "US Federal GILS" or "U.S. Federal GILS" in the Controlled Vocabulary-Local Subject Index-Local Subject Term subelement—a state presumed to indicate record-creators' intention of identifying the record as a "Core record" as delineated in the NARA *Guidelines*. No further operationalization of the "Federal Core" was achieved in this evaluation. The "Core subset" comprised 50% of the total sample.

Table C4-4
Information Object Semantics

Object	Operational Definition	Examples
Administrative Catalog	A locator listing of procedural actions related to the conduct of agency business	FERC's "Directory Of External Information Collection Requirements" PBGC's "Log Of Benefit Termination Plans" USPS's "Index Of Final Opinions And Orders"
Agency Homepage	Information mounted on an HTTP server	"Superintendent of Documents Home Page on the World Wide Web"
Bibliographic Database	An automated information system comprising metadata about bibliographic entities/publications	DOE's "OpenNet" "HUD USER"
Form	A document designed to elicit and transmit specific information from the user to the supplier, respectively	"Request for Registration for Political Risk Insurance "SSA-1710"
Job Line	A telephonic recording of employment opportunities	"DOI Employment Center"
Miscellaneous Documents In *Ad Hoc* Collection	Plurality of documents grouped by function or subject	bulletins and memoranda press releases public comments under-described "technical documents" and "reports" update notices letters speeches records
Organization	A set of human resources defined by an agency to provide specific products or services	information center/library research consortium NASA's "Flight Dynamics Facility"
Program	A prescribed set of activities and functions performed to accomplish an objective	report management records management
Publication	Discrete monographic document published one-time or in serial mode to disseminate information	annual report user's manual "The Federal Register" Regulations CD-ROM fact-sheet series procedures manual
Publications Catalog	A fixed, flat (non-machine-searchable) listing of selected or all agency publications	FEMA's "Publications Catalog"
Subject Matter Database	Single, stand-alone automated information system comprising data, records, or multiple documents on technical or administrative subject(s) and/or definable reference themes	Privacy-Act records health risks aviation accidents red cockaded woodpecker
System Of Systems	Macro-AIS comprising or integrating multiple databases and/or single-AISs	DOD's "Enterprise Information System" EPA's "Information Systems Inventory"

6.0. METHOD LIMITATIONS AND RECOMMENDATIONS TO FUTURE RESEARCHERS

The primary limitation to the procedures described for analyzing GILS record content is generalizability—the extent to which results can be assumed valid for the entire population of GILS records. The sample was small, less than 2% of the estimated total of approximately 5,000, and the sampling technique was largely convenience-driven due to time constraints. In addition, the method as employed did not provide data concerning differences in record quality among or within agencies' GILS, which might prove useful in estimating the scope of effort required in modifying elements or standardizing the characteristics of element values.

The record content analysis was extremely time-consuming, both in terms of defining mutually exclusive codes for content description and data collection. As noted above, even this small sample involved recognition of presence or absence of thousands of instances of metadata elements as well as examination and description of their values. Much of the labor burden of the current procedure could be alleviated by machine processing—e.g., for element counts, incidence of hypertext, etc. In addition, it is anticipated that the exploratory method described herein will be refined and adapted during subsequent applications, both for assessing the responsiveness of government-wide quality standards for GILS (*vis a vis* the NARA *Guidelines*) and, at the agency level, the quality of GILS record collections.

7.0. CONCLUSION

In summary, the method employed to analyze the content of GILS records proved highly satisfactory in rendering the type of results that would inform the overall evaluation. By providing a bird's-eye view of the "product on the shelf" at a given point in time, this method allows a comparison of planned vs. actual outcomes for quality. Agencies' continuous analysis and reporting of record content will serve well in complementing evaluations of the effectiveness of the NARA *Guidelines*, implementation maturity, and user satisfaction.